IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 October 2022, accepted 18 November 2022, date of publication 28 November 2022,
date of current version 1 December 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3224942

RESEARCH ARTICLE

An Improved AP Clustering Algorithm Based Critical Nodes Identification for Distribution Network With High PV Penetration

JIAWEI WU, FENG WU, (Member, IEEE), KEMAN LIN, (Member, IEEE), ZIZHAO WANG, LINJUN SHI, (Member, IEEE), AND YANG LI
College of Energy and Electrical Engineering, Hohai University, Nanjing 211100, China

Corresponding author: Keman Lin (linkeman@hhu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 52061635102.

ABSTRACT The power flow of the distributed network (DN) with high penetration of distributed PV changes significantly, such as bidirectional power flow and larger deviation of voltage magnitude, under the variation of PV power output, and the number and the position of critical nodes would change accordingly. Hence, a data-driven nodes clustering and critical node identification method is proposed in this paper. Since the relationship between different nodes are nonlinear, the autoencoder method is firstly applied to obtain the unified features of the nodes connected with different number of branches. Using the unified features, the Frechet distance is calculated to demonstrate the similarity between any two nodes. The improved affinity propagation (AP) clustering algorithm is also proposed to classify the nodes into different clusters, and the number the clusters and the critical nodes in different clusters could be obtained automatically. The electrical distance is considered in the improved AP clustering algorithm, so that nodes in different branches with similar features are avoided to be classified into one cluster. The simulations are carried out on the IEEE 123-bus system and an actual DN system, the effectiveness and efficiency of proposed critical node identification method are verified.

INDEX TERMS Critical node identification, nonlinear feature extraction, comprehensive similarity, clustering.

I. INTRODUCTION

Recently, the distributed network (DN) is integrated with high penetration of distributed PV. The intermittency and randomness of the PVs bring significant challenges to operations of the distribution system [1], [2], such as bidirectional power flow and larger voltage magnitude deviation. In this situation, the critical nodes of DN change frequently with operation of the DN, and identification of critical nodes is very important for the implementation of the preventive control strategies [3], [4], [5], [6] to maintain the stable operation of DN.

Previously, the critical nodes identification method for the power grid is mainly based on the complex network theory (CNT) [7], such as Electric betweenness [8], power flow entropy [9], singular value entropy [10], and small-world network theory [11]. Based on CNT, the undirected network model based on impedance and admittance of transmission line is applied to identify the critical nodes [7]. To consider effect of the direction of power flow, a directed network model is proposed in [12], and accuracy of the critical node identification was improved. An index for the critical node identification considering dynamics of power flow is also established in [13], so that the critical nodes can be updated when the direction of the power flow changes. Considering the topology and characteristics difference between the distribution network and transmission network, a comprehensive index including the security, economy and structure stability is established for critical node identification of the active DN in [14], and the effect of load and PV power injection changes is considered. However, it needs to be pointed out

The associate editor coordinating the review of this manuscript and approving it for publication was Shuo Sun.

that methods mentioned above are mainly based on the topology theory, and the accuracy of the critical node identification depends on the parameters of the power network strongly. Additionally, because the indexes used in previous method are calculated to present the feature of the node itself, the description of relationship between different nodes is absent, and their interaction effects could not be evaluated.

With the development of automation of the DN, its operating data are measured and collected by distribution management system (DMS). Based on the data in DMS, the operation characteristics of the DN can be analyzed using data-driven method. To identify the critical node in DN, the feature of different node and their relationships need to be extracted firstly. At this moment, the principal component analysis (PCA) is usually employed in the feature extraction of nodes in power grids [15], [16]. Since the PCA are suitable for the linear system, the improved PCA, namely kernel PCA [17], Nsytrom PCA [18], recurrence quantification analysis [19], are proposed to deal with the nonlinearity of the power grid [20], [21], [22]. Whereas nonlinear relationship between different nodes hasn't been described well using the improved PCA, and the nonlinear feature extraction method for nodes in DN need to be further investigated. Based on the extracted features of the nodes, the identification method needs to be applied to obtain the critical nodes. In traditional method, the number of the critical nodes, n, needs to be decided firstly, and n nodes with higher index are selected as the critical nodes. However, with variation the PV power output and load, the number of the critical node changes. Hence, the critical node identification method adapting the variation of the operating status of DN needs to be studied.

In this paper, a data-driven method is proposed for critical nodes identification of DN with high penetration of PV. Main contributions are as follows:

(1) The autoencoder method is applied to extract the nonlinear feature of nodes in DN, and the operation variables of nodes connected with different number of branches are aggregated to the unified features. The Frechet distance between different features are calculated as the similarity, which demonstrates the relationship between different nodes.

(2) Based on the obtained similarity values, the improved affinity propagation (AP) clustering algorithm considering electrical distance is proposed to classify the nodes, so that the nodes in different branches with similar electrical features are avoided to be classified into same cluster.

(3) Employing the density based AP clustering algorithm, the number of clusters and the cluster centers (the critical nodes) can be determine automatically, which makes that the critical node identification method proposed in this paper can meet the requirements of real-time application.

This paper is organized as follows: The model of DN is established in Section II and the structure of critical node identification proposed in this paper is presented in Section III. Section IV provides the feature extraction using autoencoder model and similarity between nodes calculated by Frechet distance. Then the improved AP clustering based critical node identification is presented in Section V. Finally, the simulation is carried on the IEEE 123-bus system and an actual DN system to verify the effectiveness of the proposed method in Section VI, followed by conclusions made in Section VII.

II. STRATEGY OF CRITICAL NODES IDENTIFICATION METHOD

The operation state of the distribution network varies under the power injection of PV, and the critical nodes are also change continuously. Based on the collected data, such as voltage amplitude, phase angle, active power and reactive power of branches, a data-driven method is proposed for the critical node identification for DN, and the strategy of the method is shown in Fig. 1. The steps of the proposed method are as follows:

Step 1: Feature extraction, the autoencoder model is applied to extract the nonlinear feature of nodes in DN.

Step 2: Similarity calculation, the Frechet distance between features of different nodes are calculated as the similarity.

Step 3: Nodes clustering, based on the obtained similarity values, the improved AP clustering algorithm is applied to classify the nodes into clusters automatically.

Step 4: Critical node identification, after nodes classification, the cluster center is identified for each cluster, which is taken as the critical node.

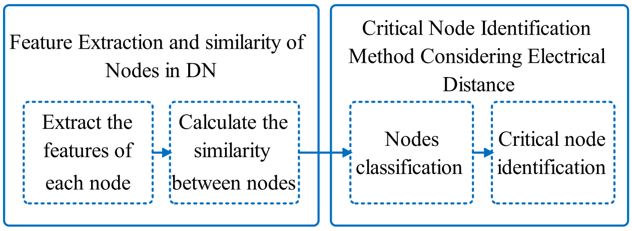

Feature Extraction and similarity of Nodes in DN
Extract the features of each node
Calculate the similarity between nodes
Critical Node Identification Method Considering Electrical Distance
Nodes classification
Critical node identification

FIGURE 1. Flow chart of critical node identification.

III. FEATURE EXTRACTION AND SIMILARITY OF NODES IN DN

An autoencoder is used to extract the time series of features of nodes under the variation of PV power output, where the system topology and the node state variables are considered, and they are extracted to the unified feature vector. Based on the unified features of nodes, Frechet distance is used to calculate the similarity between nodes to obtain the node similarity matrix.

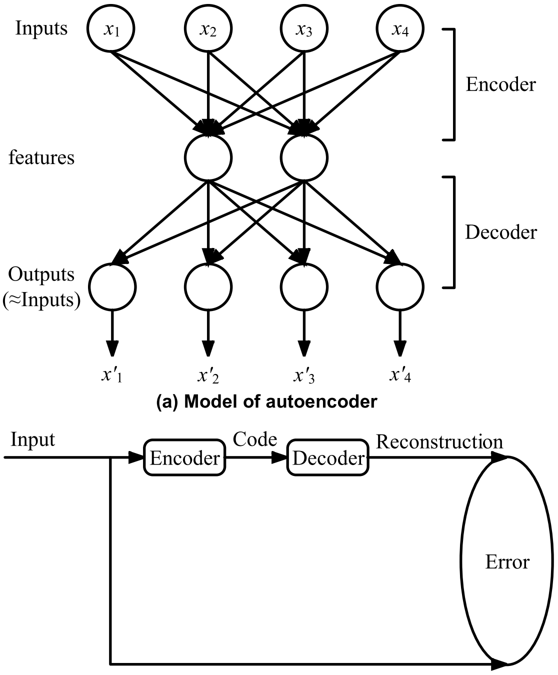
A. FEATURE EXTRACTION BASED ON NODE STATE VARIABLES

It is assumed that the state variable of node i at tth time interval is

$$x_i^t = \left[v_i^t, \theta_i^t, p_i^t, q_i^t, p_{ij}^t, q_{ij}^t \right] \tag{1}$$

where, n is the number of nodes in the DN and $v_i^t, \theta_i^t, p_i^t, q_i^t, p_{ij}^t, q_{ij}^t$ are the voltage amplitude, phase angle, injected active and reactive power of node i, active and reactive power of branch ij, respectively.

To deal with the nonlinearity of the DN, an autoencoder is employed in this paper for feature extraction of the DN. The autoencoder is a neural network based on encoder–decoder framework, which is an unsupervised learning algorithm. The encoder maps the input samples x to the feature space (i.e., encoding process), and then the decoder maps the features back to the original space to obtain the reconstructed sample x' (i.e., decoding process). The optimization target is to optimize both encoder and decoder by minimizing the reconstruction error, so as to learn the features of the sample input x.



(a) Model of autoencoder

(b) Process of autoencoder

FIGURE 2. Framework of autoencoder.

The encoding process can be represented as

$$h_i^t = \sigma\left(\omega x_i^t + b\right) \tag{2}$$

where, ω and b are the weight and deviation, respectively, x_i^t is the state variable of node i at the moment t, and σ is the activation function, which is selected as Sigmoid function in this paper.

The decoding process can be represented as

$$x_i^{\prime t} = \sigma\left(\omega^T h_i^t + c\right) \tag{3}$$

where, $x_i^{\prime t}$ is the decoded data, ω^T is the transpose of weight ω in the coding process, and c is the deviation.

The loss function L_r is trained to minimize the mean square error, so that the reconstructed state variable $x_i^{\prime t}$ is the closest to the original input x_i^t. The loss function is represented as

$$L_r = \frac{1}{m}\sum\nolimits_{k=1}^{m}\left\|x_k - x_k'\right\|_2^2 \tag{4}$$

The gradient descent method is adopted to obtain the optimal autoencoder network parameters. Inputting the state variables into the encoder, the feature of ith node at tth time interval, h_i^t, can be obtained as shown in Fig. 2.

B. CALCULATION OF SIMILARITY MATRIX

Based on the series of features of node i and j from tth time interval to $(t+n)$th time interval, two trajectories h_i and h_j can be obtained during this period. Frechet distance is employed to calculate the similarity between the feature of node i and j, because the location and temporal order of different trajectories can be considered simultaneously as shown in Fig. 3.

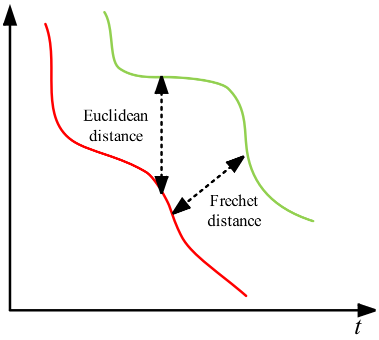


FIGURE 3. Difference between Frechet distance and Euclidean distance.

The Frechet distance of trajectories h_i^t and h_j^t are defined as reparametrized α and β within the range of [0, 1]. The infimum of the maximum value of the distance between all reparametrized $h_i\left(\alpha\left(t_w\right)\right)$ and $h_j\left(\beta\left(t_w\right)\right)$ on $t_w \in \hat{E}\,[0,1]$, and it is represented as

$$f_d\left(h_i, h_j\right) = \underset{\alpha,\beta t_w \in [0,1]}{inf}\left\{d\left(h_i\left(\alpha\left(t_w\right)\right), h_j\left(\beta\left(t_w\right)\right)\right)\right\} \tag{5}$$

where $d\left(\cdot\right)$ is the Euclidean distance, which is

$$d\left(h_i, h_j\right) = \sqrt{\sum_{t=1}^{T}\left(h_i^t - h_j^t\right)^2} \tag{6}$$

The smaller the Frechet distance is, the more similar the two trajectories are. Writing the Frechet distance between different node state variable trajectories into a matrix, the node similarity matrix can be obtained as follows:

$$F_d = \begin{bmatrix} f_d\left(h_1, h_1\right) & f_d\left(h_1, h_2\right) & \cdots & f_d\left(h_1, h_n\right) \\ f_d\left(h_2, h_1\right) & f_d\left(h_2, h_2\right) & \cdots & f_d\left(h_2, h_n\right) \\ \vdots & \vdots & \ddots & \vdots \\ f_d\left(h_n, h_1\right) & f_d\left(h_n, h_2\right) & \cdots & f_d\left(h_n, h_n\right) \end{bmatrix} \tag{7}$$

IV. CRITICAL NODE IDENTIFICATION METHOD CONSIDERING THE ELECTRICAL DISTANCE

Once the node similarity matrix is obtained, a clustering algorithm can be used to cluster nodes and the critical nodes in different cluster can be identified. Since the AP clustering is based on the "information transfer" between data points, it is not constrained by the shape of samples. Hence, the AP clustering algorithm is used to classify the nodes in DN, and electrical distance is also considered to improve the accuracy of classification. The center of the cluster is the critical node.

A. IMPROVED AP CLUSTERING ALGORITHM BASED NODE CLUSTERING

According to the principle of AP clustering, all data points are treated as potential clustering centers (called exemplars). Then, the network (similarity matrix) is formed by making connection between two data points. Finally, the clustering centers of each sample are calculated through the messages (responsibility and availability information) of edges in the network. Using similarity matrix F_d in (7), four important parameters of AP clustering can be calculated, namely the similarity $f_d\left(h_i, h_j\right)$, reference $f_d\left(h_i, h_i\right)$, responsibility information $r\left(h_i, h_k\right)$, and availability information $a\left(h_i, h_k\right)$. The relationships of the above parameters are shown in Fig. 4. The responsibility and availability information for the tth iteration are formulated respectively:

$$r_t\left(h_i, h_j\right) = f_d\left(h_i, h_j\right) - \max_{j'(j'\neq)}\left\{a_{t-1}\left(h_i, h_{j'}\right) + r_{t-1}\left(h_i, h_{j'}\right)\right\} \tag{8}$$

$$a_t\left(h_i, h_j\right) = \min\left(0, r_t\left(h_j, h_j\right) + \sum\nolimits_{i'(i'\neq i,j)} \max\left\{r_t\left(h_{i'}, h_j\right), 0\right\}\right) \tag{9}$$

$$a_t\left(h_j, h_j\right) = \sum\nolimits_{i'(i'\neq i,j)} \max\left\{r_t\left(h_{i'}, h_j\right), 0\right\} \tag{10}$$

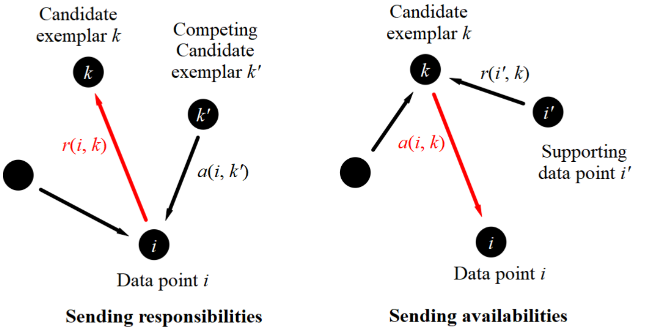


FIGURE 4. Responsibility and availability information exchange process.

The damping coefficient λ is required in the iterative process to adjust the convergence speed and attenuate numerical oscillations. λ is usually in the range of [0.5, 1], and the responsibility and availability information matrices are updated according to λ as follows:

$$r_t\left(h_i, h_j\right) = \lambda * r_{t-1}\left(h_i, h_j\right) + (1-\lambda) * r_t\left(h_i, h_j\right) \tag{11}$$

$$a_t\left(h_i, h_j\right) = \lambda * a_{t-1}\left(h_i, h_j\right) + (1-\lambda) * a_t\left(h_i, h_j\right) \tag{12}$$

The responsibility information (r_t) and availability information (a_t) are constrained to each other, decision is made according to r_t and a_t to obtain the result of clustering. The detailed steps of AP clustering can be found in literature [23].

If AP clustering algorithm is performed only with state variables, the nodes, having similar features and locating in different area, might be classified into the same cluster. Taking the IEEE 10-bus system as example, the nodes are clustered using AP as shown in Fig. 5. It can be seen that node 6 and node 9 in different area are classified into one cluster (in same color), which is obviously unreasonable.

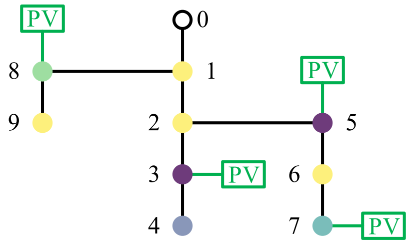


FIGURE 5. AP clustering results at the t=600 min.

In order to deal with this problem, an improved AP clustering algorithm is proposed by introducing electrical distance into similarity calculation.

Since the structure of DN are normal radial as shown in Fig. 6, it can be abstracted as a weighted radial network model. In the weighted radial network, PVs and loads are abstracted to nodes, and branches are abstracted as edges. The electrical distance between different nodes is normalized using the ratio of impedance of each branch to the maximum impedance. Thus, normalized electrical distance between node i and its neighboring node j are

$$w_{ij} = \frac{Z_{i,j}}{Z_{max}} \tag{13}$$

where $Z_{i,j}$ is the impedance between node i and j, and Z_{max} is the maximum impedance in the DN. When node i and node j are not adjacent with each other, the shortest path is taken as the electrical distance as follows:

$$w_{ij} = \sum\nolimits_{k=0}^{j-i-1} \frac{Z_{i+k,i+k+1}}{Z_{max}} \tag{14}$$

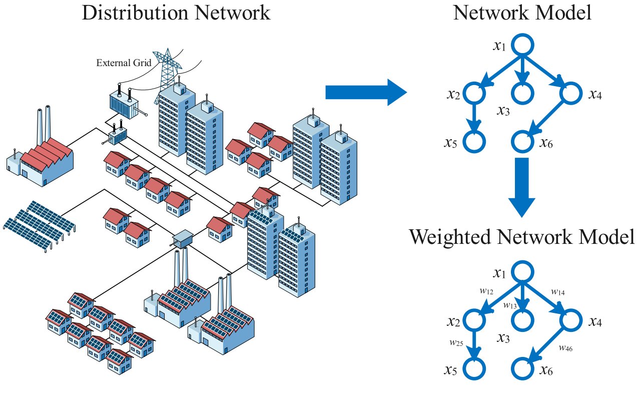


FIGURE 6. Modelling of DN.

B. CRITICAL NODE IDENTIFICATION

Based on (13) and (14), the electrical distance weight matrix can be obtained as follows:

$$W = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ w_{n1} & w_{n2} & \cdots & w_{nn} \end{bmatrix} \tag{15}$$

where, w_{nn} is the weights between node i and node j. Summing the similarity matrix F_d with the electrical distance matrix, the improved similarity matrix is obtained.

$$F_{d\_e} = (1-\gamma)\cdot F_d + \gamma \cdot W \tag{16}$$

where the coefficient γ takes value between [0,1], the larger the γ is, the greater the effect of electrical distance. The similarity matrix is restricted by the introduced electrical distance to avoid distant nodes being grouped into one cluster. In this way, the case of node 2 and 9 is avoided.

The area of each cluster reflects the influence of the cluster center. In other words, the cluster center is the node having the closest relationship with other nodes [24]. When the state variables of the cluster center change, the neighboring nodes also change accordingly. An advantage of AP clustering is that cluster centers are obtained directly, rather than indirectly through virtual cluster centers. Thus, the cluster center of each cluster after AP clustering is the critical node, as shown in Fig. 7.

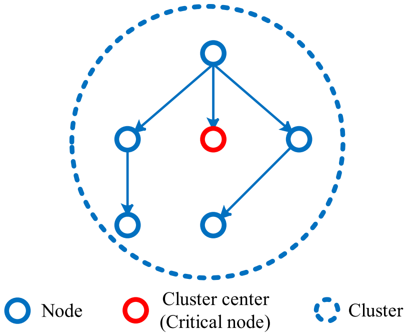


FIGURE 7. Cluster center in cluster.

The flow chart of critical node identification based on improved AP clustering algorithm proposed in this paper is shown in Fig. 8.

V. CASE STUDY

In this work, IEEE 123-bus system is taken as an example to verify the validity of the proposed method in the critical node identification for DN. Then the actual DN system in a certain area is taken to carry out a comparison between the traditional method and the method proposed in this paper.

The complex network toolbox of MATLAB is used to draw the network topology, the state variable of each node during one day is obtained by power flow calculation, with the time interval of 1 minute, and the error of ±0.5% is superimposed on the rated value. In addition, both two case studies are under static conditions.

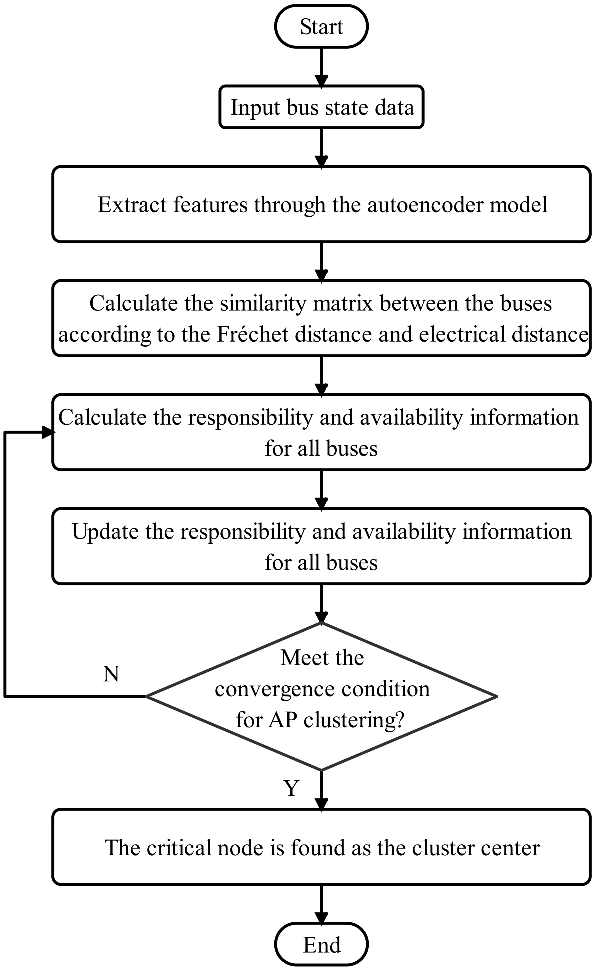

Start
Input bus state data
Extract features through the autoencoder model
Calculate the similarity matrix between the buses according to the Fréchet distance and electrical distance
Calculate the responsibility and availability information for all buses
Update the responsibility and availability information for all buses
Meet the convergence condition for AP clustering?
N
Y
The critical node is found as the cluster center
End

FIGURE 8. Flow chart of critical nodes identification based on improved AP clustering algorithm.

A. CASE 1: IEEE 123-BUS SYSTEM

The IEEE 123-bus system is used for simulation in this paper, the detailed information of IEEE 123-bus system can be found in [25]. In the system, 9 PV plants with total capacity of 3.6 MW are integrated in the system, whose capabilities and locations are listed in Table 1. The IEEE 123-bus system is shown in Fig. 9.

TABLE 1. Capacity and location of PV units.

PV location	PV capacity (MVA)
33, 42, 93	0.2
18, 64, 97	0.4
57, 81, 119	0.6

1) NODE SIMILARITY SEARCH

The parameters of the autoencoder are shown in Table 2. The input data includes node voltage magnitude, phase angle, active and reactive power, and corresponding branch active and reactive power. The number of features needs to be selected manually in the autoencoder, and the relationship

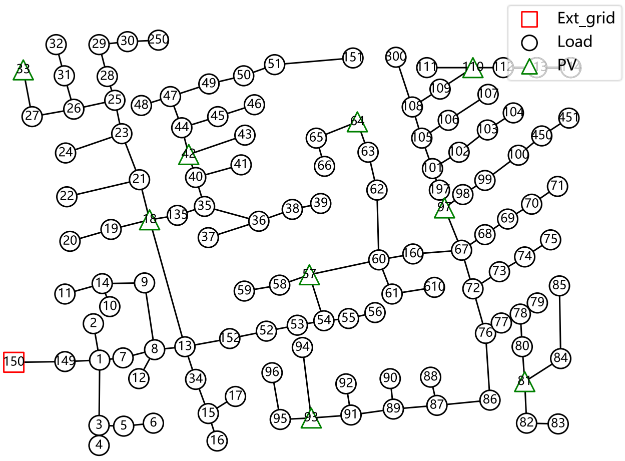


FIGURE 9. Diagram of IEEE 123-bus system.

between the dimensions retained after feature extraction should be as small as possible [26]. In this paper, the dimensionality is set to 3.

TABLE 2. Autoencoder parameter configuration.

Number of hidden layers	Number of neurons per layer	Number of features	Learning rate	Number of training times
3	64	3	0.001	2,000

As can be seen from Fig. 10, the data after the autoencoder model reduction is almost identical to the origin data. Therefore, the autoencoder is suitable for data processing with strong nonlinear characteristics such as DN with high penetrated PVs.

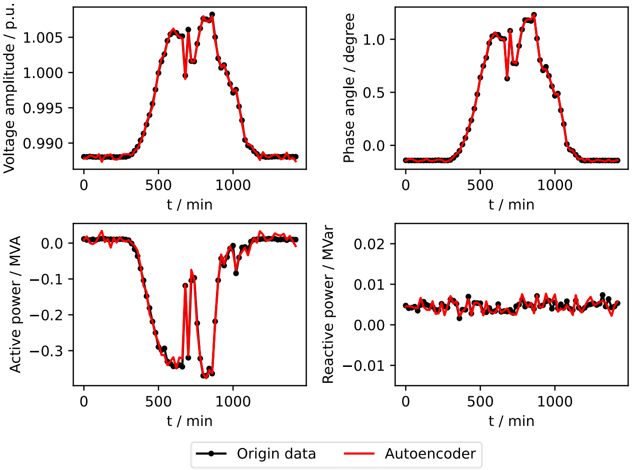


FIGURE 10. Autoencoder model for state variables of PV node 33.

After feature extraction, the similarity index can be calculated through Frechet distance. F_d is calculated by the conventional AP clustering and F_{d_e} in (16) is calculated by the improved AP clustering proposed in this paper, when γ takes a value of 0, F_{d_e} degenerates to F_d. In this case, node 13 is selected (with four neighboring nodes: 8, 18, 34, and 152) to compare the similarity matrix F_d and the improved similarity matrix F_{d_e}, the results are shown in Fig. 11. The smaller the similarity index between two nodes, the more likely these nodes will be clustered into the same cluster. The similarity matrix F_d of node 13 depends only on the extracted features. The four nodes with the lowest similarity index value to node 13 (node 16, 42, 51 and 250) are not topologically neighboring nodes of node 13. Therefore, the clustering result of conventional AP clustering is not in consistent with the practical DN topology. In contrast, the improved similarity matrix F_{d_e} introduces electrical distance, and the four nodes with the lowest similarity index value to node 13 (node 8, 18, 34 and 152) are neighboring nodes of this node. After the test, a γ value of 0.3 can effectively combine the impact of F_d and W. The clustering result of improved AP clustering is in consistent with the practical DN topology.

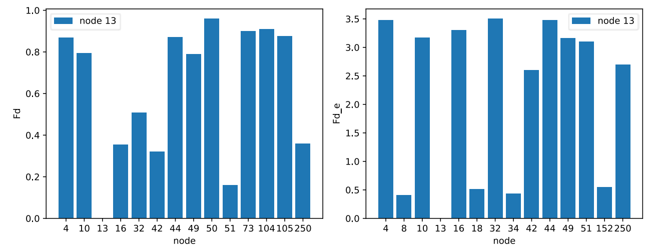


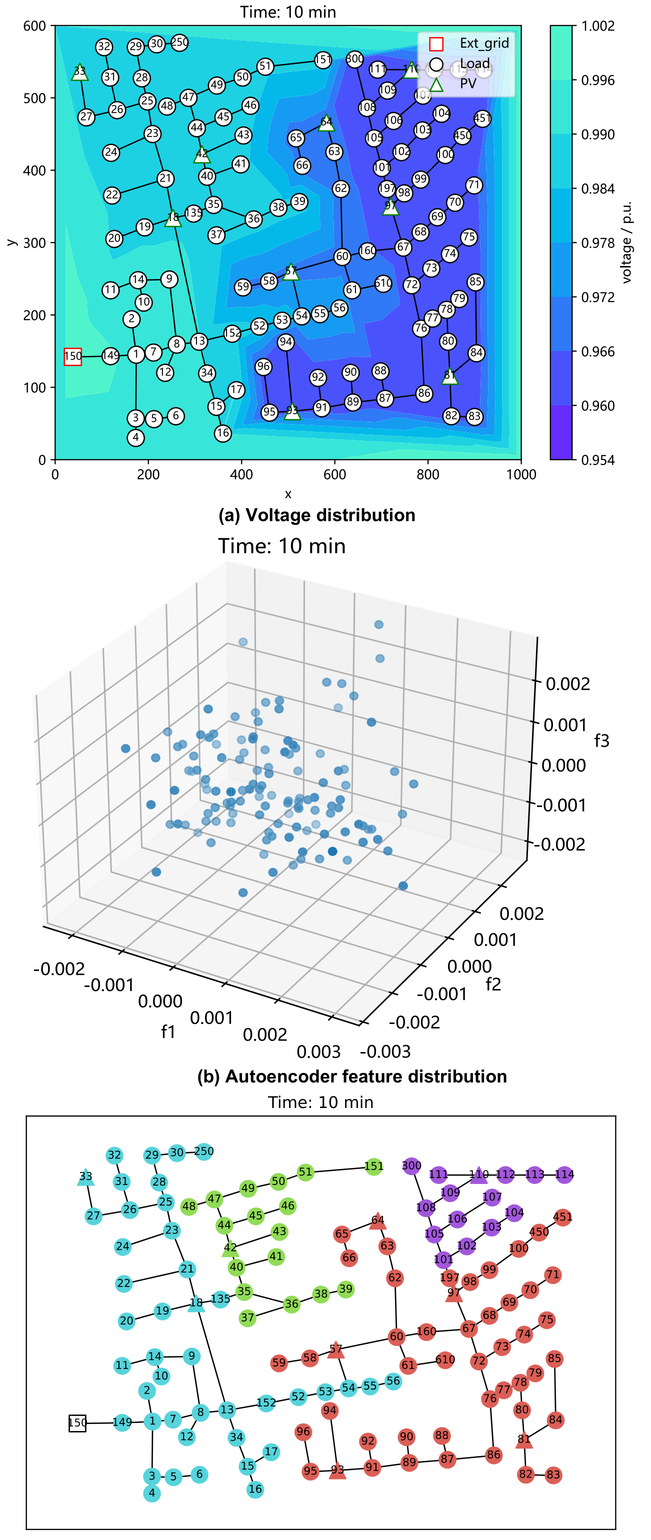
FIGURE 11. F_d and F_{d_e} of node 13.

2) CLUSTERING AND CRITICAL NODE IDENTIFICATION

After obtaining the similarity matrix, the critical node identification can be performed. t = 10 min (early morning with little PV power output) and t = 800 min (midday with high PV power output) are selected for comparative analysis. The clustering results for the two cases are shown in Fig. 12 and 13, respectively.

At t = 10 min, PVs generate low power. The nodes near the root node have higher voltage magnitude, while the nodes far from the root node have lower voltage magnitude as shown in Fig. 12(a). The feature distribution after extraction is concentrated and the distances between feature points are less than 0.1, as shown in Fig. 12(b). The value in similarity matrix F_{d_e} is small, generally less than 0.004. (16) indicates that the difference in similarity between nodes at this time is mainly reflected in the electrical distance, so the electrical distance played a great role in the clustering process. The number of clusters after clustering is smaller, and the clustering results are mostly dominated by the root nodes, as shown in Fig. 12(c).

At t = 800 min, PVs generate high power. The node voltage magnitude in the network is higher near the PV access nodes and lower away from the PV access nodes, as shown in Fig. 13(a). The feature distribution after extraction is dispersed and the distances between feature points are more than

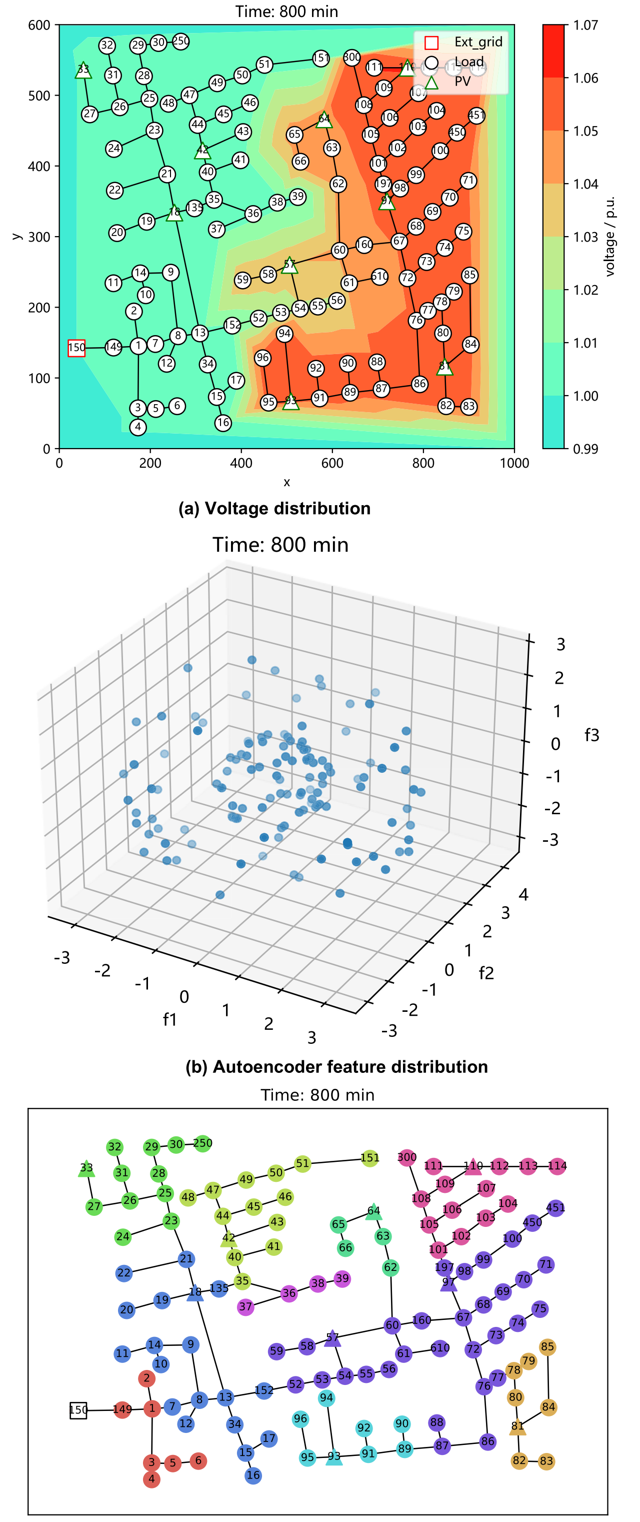


(a) Voltage distribution

(b) Autoencoder feature distribution

(c) Clustering results

FIGURE 12. Voltage distribution, feature distribution, and clustering results at t = 10 min.



(a) Voltage distribution

(b) Autoencoder feature distribution

(c) Clustering results

FIGURE 13. Voltage distribution, feature distribution, and clustering results at t = 800 min.

2 as shown in Fig. 13(b). The value of the resulting similarity matrix F_{d_e} is large, with a maximum value close to 3.5. (16) indicates that the electrical distance and F_{d_e} function jointly at this moment. The number of clusters is higher after clustering, and the clustering results are mostly dominated by the PV access nodes, as shown in Fig. 13(c).

The cluster center is essentially a critical node, cluster centers and their neighboring nodes at $t = 10$ min and $t = 800$ min are shown in Table 3. At $t = 10$ min, PVs generate low power, electrical distance plays the major role, so the critical nodes are the root nodes close to the external grid. At $t = 800$ min, PVs generate high power, the neighboring nodes are influenced by the PV power output, resulting in higher voltage magnitude. F_d makes the PV node more likely to become the cluster center, and the electrical distance makes the root node more likely to become the cluster center, F_d and electrical distance constrain each other. Therefore, the critical nodes are nodes near the PV access nodes.

TABLE 3. The cluster center and its neighboring nodes and the most critical nodes at t=10 min and t=800 min.

Time (min)	Number of clusters	critical nodes
10	4	149, 60, 35, 101
800	10	1, 13, 26, 40, 36, 91, 160, 81, 108, 63

In this paper, data from a typical day with a 20-minute interval are selected for simulation to observe the effect of PV power output on the clustering results and the critical node identification. When the PVs power generation increases, the nodes near the PV access node are affected by PV power output, resulting in higher voltage magnitude, while the nodes far from the PV do not change much. Therefore, the number of clusters increased as the PV power output increased, as shown in Fig. 14.

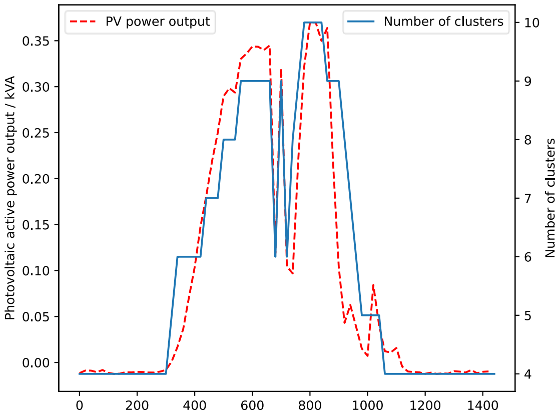


FIGURE 14. PV power output and number of clusters at separate times.

Fig. 15 shows the change process of critical node at different time in a region of the IEEE 123-bus system. As the PVs power output increases, the influence of PV power output on voltage of neighboring nodes increases gradually, so the critical nodes will normally move from root nodes to the nodes before the PV nodes.

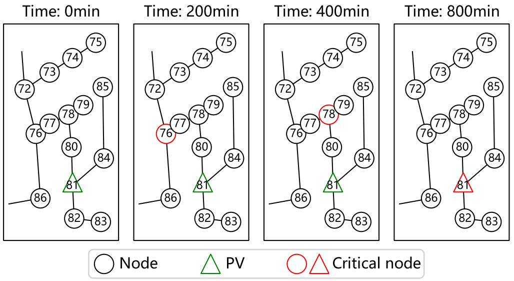


FIGURE 15. Process of critical node change at different times.

B. CASE 2: AN ACTUAL DN IN A CERTAIN AREA

An actual DN system in a city in Jiangsu, China, is used to carry out a comparison between the traditional method and the method proposed in this paper to verify the effectiveness and superiority of the proposed method. The system has over 300 nodes connected to a total of 16 PV plants, as shown in Fig. 16.

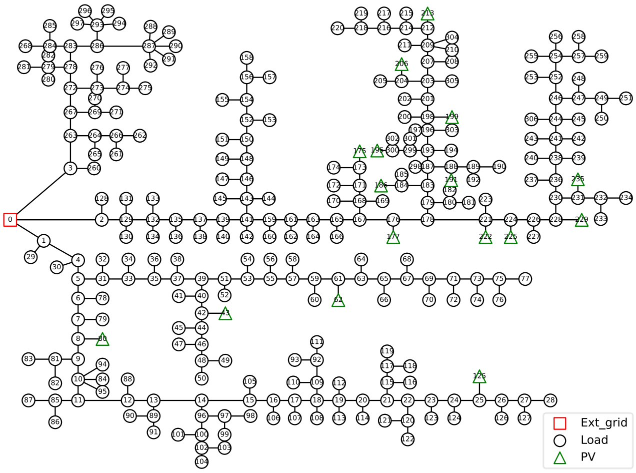


FIGURE 16. Diagram of an actual DN in a certain area.

Comparison of the method based on the combination of CNT and node resilience [14] is carried out. The results are shown in Table 4 (t=10 min) and 5 (t=800 min).

From Table 4, the PVs generate low power, it can be seen that the results obtained by proposed method are almost the same with those obtained by [14]. From Table 5, the PVs generate high power, the results obtained in this paper are slightly different from those of [14], but the results are generally close. So, the proposed method is proved to be credible and effective.

TABLE 4. Critical node identification on different methods at t=10 min.

Method in Literature [14]	Method in this paper
2	2
3	3
5	4
10	10
167	167
224	226

By comparing the results of the two tables, it can be found that the method in [14] focuses on the structural stability of

TABLE 5. Critical node identification on different methods at t=800 min.

Method in Literature [14]	Method in this paper
3	3
6	8
10	10
20	25
51	42
57	62
167	179
201	203
226	231

the network, and the results of the critical node identification still show structure nature in case of high PV power output. In contrast, the method in this paper considers both the effect of PV power output and the structural characteristics of the network, so that the critical nodes show a proximity to the PV nodes when PV power output is high. In addition, traditional methods like [14] create indicators for each node and require artificially specified thresholds to determine the number of critical nodes. The method proposed in this paper is based on the AP clustering algorithm, which does not require an artificially specified number of critical nodes and is therefore more suitable for large scale DNs.

The computing efficiency for critical node identification between two cases are listed in Table 6, it can be seen that the time consumption of the method in this paper is higher than that of the traditional method in [14], because of the higher computation burden of the big data analysis method. It also needs to be mentioned that it takes only 2 seconds to identify the critical nodes for the DN with more than 300 nodes, which is quick enough for regular voltage control of DN.

TABLE 6. The time consuming of two cases.

Method	IEEE 123-bus system (s)	Actual DN system (s)
Method in this paper	1.2741	2.1096
Method in literature [14]	0.6239	1.0981

VI. CONCLUSION

In this paper, A data-driven critical node identification method for the high penetrated PV DN has been proposed. Compared to the traditional topology based methods, the proposed method has the following advantages:

1) Since both node similarity and electrical distance have been considered in the improved AP clustering algorithm, the node in DN can be classified correctly and efficiently.
2) Using the clustering based methods, the node in DN can be classified into clusters automatically, and clusters centers is identified as the critical node. The number of critical nodes varies with the injection power of PV.
3) With the increasing power output of PV, the positions of critical nodes also change, which normally move from the root nodes to the nodes before the PV nodes.

The identification of critical node can provide supporting information for preventive control and optimal scheduling to meet the requirements of real-time control. The further research will be carried to study the optimal operation and control strategies based on the identified critical nodes for DN containing large amounts of distributed PV to ensure safe and reliable operation of the DN.

REFERENCES

[1] N. C. Scott, D. J. Atkinson, and J. E. Morrell, "Use of load control to regulate voltage on distribution networks with embedded generation," *IEEE Trans. Power Syst.*, vol. 17, no. 2, pp. 510–515, May 2002.

[2] R. Caire, N. Retiere, S. Martino, C. Andrieu, and N. Hadjsaid, "Impact assessment of LV distributed generation on MV distribution network," in *Proc. IEEE Power Eng. Soc. Summer Meeting*, Chicago, IL, USA, Jul. 2002, pp. 1423–1428.

[3] L. Ran, J. Baoyuan, Z. Fan, F. Hang, and D. Qingang, "Research on identification of key nodes in power system based on spectral graph theory," *Power Syst. Protection Control*, vol. 46, no. 11, pp. 14–22, 2018.

[4] S. Wang, W. Lv, J. Zhang, S. Luan, C. Chen, and X. Gu, "Method of power network critical nodes identification and robustness enhancement based on a cooperative framework," *Rel. Eng. Syst. Saf.*, vol. 207, Mar. 2021, Art. no. 107313.

[5] J. Xu, W. Lv, J. Zhang, S. Luan, C. Chen, and X. Gu, "Identification of power grid key parts based on improved complex network model," *Autom. Electr. Power Syst.*, vol. 40, no. 10, pp. 53–61, 2016.

[6] J. Wang, C. Wei, X. Pan, Y. Liu, Y. Li, and W. Xie, "Fast matching method for key nodes of smart distribution network oriented to operational optimization requirements," *Power Syst. Technol.*, vol. 43, no. 3, pp. 848–855, 2019.

[7] B. Liu, Z. Li, X. Chen, Y. Huang, and X. Liu, "Recognition and vulnerability analysis of key nodes in power grid based on complex network centrality," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 65, no. 3, pp. 346–350, Mar. 2018.

[8] S. Sun, X. Li, F. Zhang, W. Shi, and C. Hao, "Identification of vulnerable lines in the distribution network based on network structure importance and potential hazard vulnerability," *Power Syst. Protection Control*, vol. 46, no. 14, pp. 107–113, 2018.

[9] F. Wenli, Z. Xuemin, M. Shengwei, H. Shaowei, W. Wei, and D. Lijie, "Vulnerable transmission line identification using ISH theory in power grids," *IET Gener., Transmiss. Distrib.*, vol. 12, no. 4, pp. 1014–1020, Feb. 2018.

[10] T. Wang, X. Yue, X. Gu, S. Zhang, and B. Zhao, "Power grid critical node identification based on singular value entropy and power flow distribution entropy," *Electr. Power Autom. Equip.*, vol. 36, no. 4, pp. 46–53, 2016.

[11] Y. Zhao, Y. An, and Q. Ai, "Research on size and location of distributed generation with vulnerable node identification in the active distribution network," *IET Gener., Transmiss. Distrib.*, vol. 8, no. 11, pp. 1801–1809, Aug. 2014.

[12] Y.-S. Li, D.-Z. Ma, H.-G. Zhang, and Q.-Y. Sun, "Critical nodes identification of power systems based on controllability of complex networks," *Appl. Sci.*, vol. 5, no. 3, pp. 622–636, Sep. 2015.

[13] D.-S. Yang, Y.-H. Sun, B.-W. Zhou, X.-T. Gao, and H.-G. Zhang, "Critical nodes identification of complex power systems based on electric cactus structure," *IEEE Syst. J.*, vol. 14, no. 3, pp. 4477–4488, Sep. 2020.

[14] W. Zhang, K. Liu, W. Sheng, S. Du, and D. Jia, "Critical node identification in active distribution network using resilience and risk theory," *IET Gener., Transmiss. Distrib.*, vol. 14, no. 14, pp. 2771–2778, Jul. 2020.

[15] X. Liu, D. M. Laverty, R. J. Best, K. Li, D. J. Morrow, and S. McLoone, "Principal component analysis of wide-area phasor measurements for islanding detection—A geometric view," *IEEE Trans. Power Del.*, vol. 30, no. 2, pp. 976–985, Apr. 2015.

[16] M. Rafferty, X. Liu, D. M. Laverty, and S. McLoone, "Real-time multiple event detection and classification using moving window PCA," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2537–2548, Sep. 2016.

[17] Z. Li, U. Kruger, L. Xie, A. Almansoori, and H. Su, "Adaptive KPCA modeling of nonlinear systems," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2364–2376, May 2015.

[18] G. Liu, H. Chen, X. Sun, N. Quan, L. Wan, and R. Chen, "Low-complexity nonlinear analysis of synchrophasor measurements for events detection and localization," *IEEE Access*, vol. 6, pp. 4982–4993, 2018.

[19] P. Bhui and N. Senroy, "Application of recurrence quantification analysis to power system dynamic studies," *IEEE Trans. Power Syst.*, vol. 31, no. 1, pp. 581–591, Jan. 2016.

[20] L. Cai, N. F. Thornhill, S. Kuenzel, and B. C. Pal, "Wide-area monitoring of power systems using principal component analysis and k-nearest neighbor analysis," *IEEE Trans. Power Syst.*, vol. 33, no. 5, pp. 4913–4923, Sep. 2018.

[21] S. Liu, Y. Zhao, Z. Lin, Y. Liu, Y. Ding, L. Yang, and S. Yi, "Data-driven event detection of power systems based on unequal-interval reduction of PMU data and local outlier factor," *IEEE Trans. Smart Grid*, vol. 11, no. 2, pp. 1630–1643, Mar. 2020.

[22] H.-F. Wang, C.-Y. Zhang, D.-Y. Lin, and B.-T. He, "An artificial intelligence based method for evaluating power grid node importance using network embedding and support vector regression," *Frontiers Inf. Technol. Electron. Eng.*, vol. 20, no. 6, pp. 816–828, Jun. 2019.

[23] M. S. Park, J. H. Na, and J. Y. Choi, "PCA-based feature extraction using class information," in *Proc. IEEE Int. Conf. Syst., Man Cybern.*, vol. 1, Oct. 2005, pp. 341–345.

[24] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, Feb. 2007.

[25] K. Wang, J. Zhang, and D. Li, "Adaptive affinity propagation clustering," *Acta Autom. Sinica*, vol. 33, no. 12, pp. 1242–1246, 2007.

[26] N. Samaan, M. A. Elizondo, B. Vyakaranam, M. R. Vallem, X. Ke, R. Huang, J. T. Holzer, S. Sridhar, Q. Nguyen, Y. V. Makarov, X. Zhu, J. Wang, and N. Lu, "Combined transmission and distribution test system to study high penetration of distributed solar generation," in *Proc. IEEE/PES Transmiss. Distrib. Conf. Expo. (TD)*, Apr. 2018, pp. 1–9.

[27] I. Jolliffe, *Principal Component Analysis*. New York, NY, USA: Wiley, 2002.

JIAWEI WU received the B.S. degree from Hohai University, Changzhou, China, in 2016. He is currently pursuing the Ph.D. degree with the College of Energy and Electrical Engineering, Hohai University.

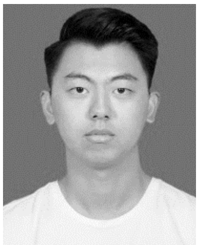
His current research interests include dynamic state estimation and situational awareness.

FENG WU (Member, IEEE) received the B.Eng. and M.Sc. degrees in electrical engineering from Hohai University, Nanjing, China, in 1998 and 2002, respectively, and the Ph.D. degree in electrical engineering from the University of Birmingham, Birmingham, U.K., in 2009.

He is currently a Professor with Hohai University. His research interests include modeling and control of the renewable energy generation.

KEMAN LIN (Member, IEEE) was born in Nanjing, China, in 1987. She received the B.S. and Ph.D. degrees in electrical engineering from Southeast University, Nanjing, in 2010 and 2016, respectively.

She is currently an Associate Professor of electrical engineering with Hohai University, Nanjing. Her current research interests include power system stability analysis and control, and modeling and control relating to renewable energy integration.



ZIZHAO WANG received the B.S. degree from Hohai University, Nanjing, China, in 2017, where he is currently pursuing the Ph.D. degree with the College of Energy and Electrical Engineering.

His current research interest includes application of new energy in power systems.

LINJUN SHI (Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from Hohai University, Nanjing, China, in 1999 and 2003, respectively, and the Ph.D. degree in electrical engineering from Southeast University, Nanjing, in 2010.

He was as a Visiting Scholar at Baylor University, Waco, TX, USA, in 2014. He is currently an Associate Professor with the College of Energy and Electrical Engineering, Hohai University. His research interests include power system analysis and control, new energy, and energy storage applications to power systems.

YANG LI received the B.E. and Ph.D. degrees in electrical engineering from Zhejiang University, Hangzhou, China, in 2014 and 2019, respectively.

He spent a year as a Visiting Ph.D. Student at the Illinois Institute of Technology. He is currently an Associate Professor with the College of Energy and Electrical Engineering, Hohai University. His main research interests include integrated energy system operation and complementary power generation systems.